ELSEVIER

Knowledge-Based Systems 15 (2002) 243–250

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

A data mining approach to discover unusual folding regions in genome sequences

Shu-Yun Le[a,*], Wei-min Liu[b], Jacob V. Maizel Jr.[a]

[a]*Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, NIH, Building 469, Room 151, Frederick, MD 21702, USA*

[b]*Department of Computer and Information Science, Indiana University, 723 W. Michigan St., Indianapolis, IN 46202-3216, USA*

Received 15 March 2001; revised 4 May 2001; accepted 31 May 2001

Abstract

Numerous experiments and analyses of RNA structures have revealed that the local distinct structure closely correlates with the biological function. In this study, we present a data mining approach to discover such unusual folding regions (UFRs) in genome sequences. Our approach is a three-step procedure. During the first step, the quality of a local structure different from a random folding in a genomic sequence is evaluated by two z-scores, significance score (SIGSCR) and stability score (STBSCR) of the local segment. The two scores are computed by sliding a fixed window stepped a base along the sequence from the start to end position. Next, based on the *non-central Student's t distribution theory* we derive a linearly transformed *non-central Student's t distribution* (LTNSTD) to describe the distribution of SIGSCR and STBSCR computed in the sequence. In the third step, we extract these significant UFRs from the sequence whose SIGSCR and/or STBSCR are greater or less than a given threshold calculated from the derived LTNSTD. Our data mining approach is successfully applied to the complete genome of *Mycoplasma genitalium* (*M. gen*) and discovers these statistical extremes in the genome. By comparisons with the two scores computed from randomly shuffled sequences of the entire *M. gen* genome, our results demonstrate that the UFRs in the *M. gen* sequence are not selected by chance. These UFRs may imply an important structure role involved in their sequence information. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Data mining; Statistical model; RNA/DNA folding; UFR

1. Introduction

Complete genomic sequence data are being accumulated at an unprecedented pace. A wide variety of computational methods for analyzing genomic sequences have been developed [1,2]. Most of the problems in these methods are essentially statistical. Computational analyses of the distinct sequence pattern can help to understand the structure and function of genomic sequences. The discovery of biological knowledge from sequence data consisting of bases A, C, G, T/U in biological databases, such as Genbank, is especially important in a post-genomics age.

RNA is a single-stranded conformationally polymorphic macromolecule with its nucleotide sequence identical to that of one of the DNA strands except for a base replacement of T to U. The RNA sequence often folds back on itself between complementary segments to form various local structures guided by Watson–Crick rules. In addition to the Watson–Crick A–U and G–C base pairs, wobble G–U base pairs also contribute to the thermodynamic stability of an RNA structure. It has been demonstrated that some structures folded by local RNA segments are functional elements of the control for gene regulations in different levels [3,4]. These functional elements are often closely associated with unusual folding regions (UFRs) where the folding free energy of the UFR is significantly lower than that expected by chance [5–12]. The development of an efficient data mining approach to extract these potentially functional structured elements in the sequence database is highly desirable.

Knowledge discovery of functional structured elements in a genomic sequence is an important step to reach our goal from genome data to biological knowledge. The thermodynamic stability of an RNA/DNA fragment in the genome is often measured by the free energy of the formation of the folded RNA/DNA segment. Based on accumulated data [3,4,13], UFRs in an RNA sequence are assessed by the two z-scores, significant score (SIGSCR) and stability score (STBSCR) [13,14]. SIGSCR signifies the difference

* Corresponding author. Tel.: +1-301-846-5576; fax: +1-301-846-5598.
E-mail address: shuyun@orleans.ncifcrf.gov (S.-Y. Le).

0950-7051/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0950-7051(01)00146-0

of thermodynamic stability between a local, natural RNA fragment and the average of its randomly shuffled sequences. Similarly, STBSCR indicates the difference of the stability between a specific fragment at a given place and the average from all other fragments of the same size in the sequence. As an example of our data mining, we analyze the complete genome sequence of *Mycoplasma genitalium* (*M. gen*).

Our data mining approach consists of three steps. In the first stage, we compute SIGSCR and STBSCR by sliding a fixed window with a step of one base along the sequence from the start to end position. Our statistical analysis shows that the distributions of the two *z*-scores in the sequence do not follow a simple normal distribution. In order to obtain useful information from an extraordinarily large number of sample observations in the analysis, we have to derive a reliable statistical model to describe the distributions of the two *z*-scores. In the second step we develop a linearly transformed *non-central Student's t* statistical model to delineate the distributions of SIGSCR and STBSCR in the entire genomic sequence by means of a *non-central Student's t distribution* theory [15]. Statistical tests show that the linearly transformed *non-central Student's t distribution* (LTNSTD) is a good statistical model to describe the distributions of the two scores computed in the genome. In the last step, the significant UFRs that are either much more stable or unstable than expected by chance are discovered based on the derived, well-fitted LTNSTD.

As a comparison, we also compute the distributions of SIGSCR and STBSCR in the randomly shuffled sequence of the complete *M. gen* genome. Our results further demonstrate that the statistical extremes of UFRs are not selected by chance in *M. gen*. The UFRs in the genome may imply the biological functions of the primary sequence data and provide useful information in further searching for functional structured elements involved in the control of regulatory genes [5–12].

2. Mathematical background

2.1. SIGSCR and STBSCR of a folding segment

The quality of a local structure in a DNA/RNA sequence is often evaluated by the thermodynamic stability of the structured segment. The greater the free energy of the formation of the structure in negative numbers, the more stable the folded structure of a fragment. In this study, the biological information of such structured fragments in an RNA sequence is evaluated by SIGSCR and STBSCR of a local segment. SIGSCR and STBSCR are a standard *z*-score and given by

$$\mathrm{SIGSCR} = (E - E_{\mathrm{r}})/std_{\mathrm{r}}$$

and

$$\mathrm{STBSCR} = (E - E_{\mathrm{w}})/std_{\mathrm{w}}$$

where E is the folded lowest free energy computed from a local segment in the sequence, E_{r} is the sample mean and std_{r} is the sample standard deviation of the lowest free energies computed from folding a large number of randomly shuffled segments of the same size and same base compositions as the local segment. Similarly, E_{w} and std_{w} are the sample mean and standard deviation of the lowest free energies obtained by folding all segments of the same size that are generated by taking successive, overlapping, fixed length segments stepped one base at a time from the start to end position of the sequence [13,14].

2.2. Linearly transformed non-central Student's t distribution

The *non-central Student's t distribution* (NSTD) is an asymmetric continuous distribution with range $(-\infty.\ +\infty)$. It has two parameters [15]: the degree of freedom, f (a positive integer); and the non-centrality parameter, δ (a real number). Its probability density function (PDF) [16] can be expressed as

$$P(x;f,\delta) = \frac{1}{2^{(f+1)/2}\Gamma(f/2)\sqrt{\pi f}} \int_0^{\infty} y^{(f-1)/2} \times \exp\left[-\frac{1}{2}y - \frac{1}{2}\left(x\sqrt{\frac{y}{f}} - \delta\right)^2 \right] \mathrm{d}y \qquad (1)$$

where x is a random variable, y is an integration variable and Γ is the well-known gamma function. Its r-th moment about the origin [15] is

$$\mu'_r = \left(\frac{f}{2}\right)^{r/2} \frac{\Gamma((f-r)/2)}{\Gamma(f/2)} \sum_{j=0}^{r/2} \binom{r}{2j} \frac{(2j)!}{2^j j!} \delta^{r-2j}, (f > r). \qquad (2)$$

To simplify the notation, we introduce

$$g(f) = \frac{\Gamma((f-1)/2)}{\Gamma(f/2)}. \qquad (3)$$

Let the observed data, SIGSCR and STBSCR be $\{y_i,\ 1 \le i \le n\}$. Consider a linear transformation:

$$y_i = ax_i + b, a > 0. \qquad (4)$$

Let x_i, $1 \le i \le n$ be distributed as an NSTD t, whose degree of freedom is f and non-centrality parameter is δ. For a given degree of freedom f, we estimate the parameters a, b and δ by assuming the sample mean, sample variance and sample coefficient of skewness (k) of variables x_i ($1 \le i \le n$) are equal to the mean, variance and coefficient of skewness of the NSTD, that are shown in the following three equations [15,23]:

$$\frac{\bar{y} - b}{a} = \sqrt{\frac{f}{2}} g(f)\delta, \qquad (5)$$

$$\frac{s_y^2}{a^2} = \frac{f}{f-2}\left[1 + \delta^2\left(1 - \frac{f-2}{2}g^2(f)\right)\right], \qquad (6)$$

Table 1
Statistics for significance score (SIGSCR) and stability score (STBSCR) computed by sliding the window of 100, 300 and 500 bp along the complete *Mycoplasma genitalium* (*M. gen*) genome (strain G37). (The random samples are selected from the corresponding, complete data of samples computed by three fixed windows of 100, 300 and 500 bases. The distance between any two observations in the random sample is equal to or greater than 100 bases. There are three samples for complete SIGSCR data and STBSCR data. There are also three random samples for the randomly selected SIGSCR and STBSCR data)

Window size (bp)	Sample		SIGSCR			STBSCR		
	Type	Size (*N*)	Mean	Std	Skewness	Mean	Std	Skewness
100	Complete	579,975	−0.599	1.124	−0.558	−0.000	1.000	−0.607
100	Random	5000	−0.597	1.127	−0.488	0.001	1.001	−0.633
300	Complete	579,775	−1.397	1.307	−0.399	−0.000	1.000	−0.748
300	Random	5000	−1.403	1.295	−0.619	−0.001	0.999	−0.761
500	Complete	579,575	−2.036	1.609	−1.162	−0.000	1.000	−0.854
500	Random	5000	−2.038	1.594	−1.170	0.003	0.999	−0.846

$$k = \sqrt{\frac{f-2}{2}}\,\frac{g(f)\delta\left[3 + (f-2)\delta^2\left((f-3)g^2(f) - \frac{2f-7}{f-2}\right)\right]}{(f-3)\left[1 + \delta^2\left(1 - \frac{f-2}{2}g^2(f)\right)\right]^{3/2}} \tag{7}$$

where $\bar{y}$ and s_y are the sample mean and sample standard deviation of variables y_i that are observations in the sample of data SIGSCR or STBSCR.

3. First step of our data mining: computing SIGSCR and STBSCR

In the first step of our data mining approach, SIGSCR and STBSCR in a sequence are computed by the program SIGSTB, a modified version of SEGFOLD [14], using fixed windows of 100, 300 and 500 bases. The program SIGSTB first computes E, E_r, std_r and SIGSCR for the fragment with the same size as the selected window from the beginning of the sequence. The lowest free energy E is computed by folding the segment using the dynamic programming algorithm [17] and Turner energy rules [18]. E_r and std_r of the fragment are calculated from the tabulated coefficients based on the segment length and its base compositions, if the percentage of base $G+C$ in the segment is less than 75% and each base percentage is larger than 3% [21]. Otherwise, the mean energy E_r and its standard deviation std_r are computed by folding 100 randomly shuffled sequences of the fragment by the dynamic programming algorithm [17]. The computation is continued by sliding the fixed window one base at a time until the window reaches the end position of the sequence. Next, E_w and std_w are calculated from those E values computed for all segments as described above. Finally, STBSCR values for each segment in the sequence are calculated.

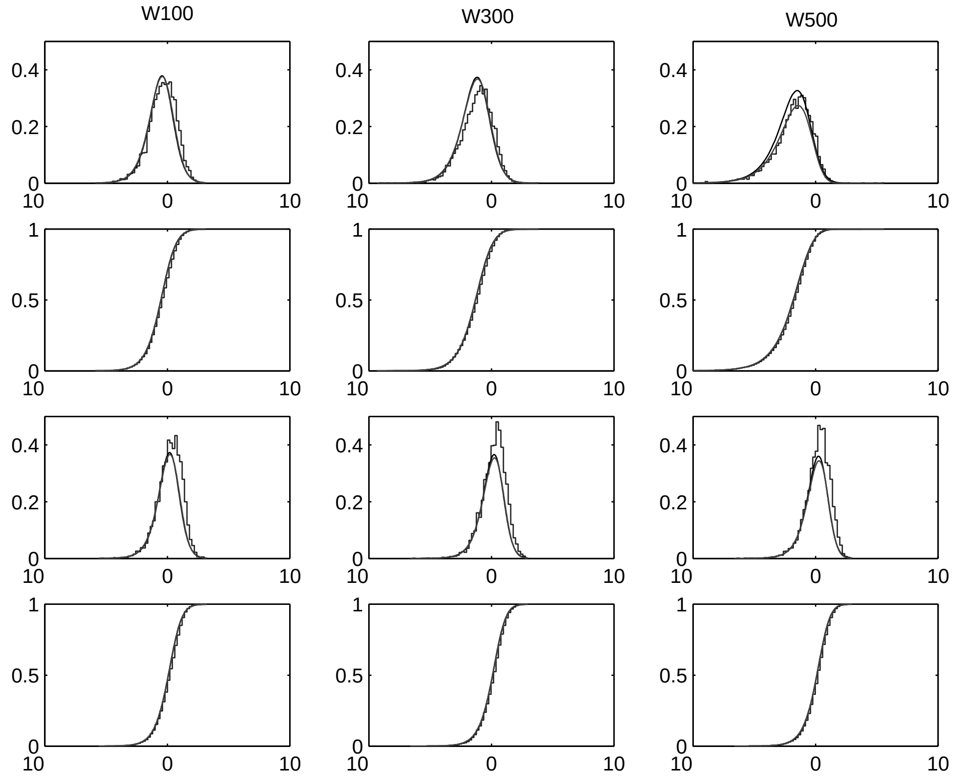
Using the three fixed windows of 100, 300 and 500 bases we collect three samples for data SIGSCR and three samples for STBSCR from *M. gen*. In the computation, the *M. gen.* sequence of 580,074 nucleotides is divided into two overlapping sequences of 1–300,600 and 300,001–580,074 because of the computing limitation of the program SIGSTB. Certainly, the lowest free energies could also be computed by other dynamic programming algorithms such as those developed by Elloumi [19,20]. It has been demonstrated from our extensive tests [21] that the computed E_r and std_r from the empirical formulas are in good agreement with those computed from a Monte Carlo simulation of 300 randomly shuffled sequences.

4. Second step of our data mining: deriving a LTNSTD for SIGSCR and STBSCR

Since neighboring scores in the six samples are possibly not fully independent, we also take a random sample with size of 5000 observations (SIGSCR or STBSCR) for each of the six samples so that the distance between any two neighboring observations in the randomly selected sample is equal to or larger than 100 bases. We compute the sample mean ($\bar{y}$), sample standard deviation (s_y) and sample coefficient of skewness (k) for data SIGSCR and STBSCR in these randomly selected samples. For a given degree of freedom f and parameter k, we solve Eq. (7) to obtain δ, then substitute the value of δ into Eq. (6) to get a and b from Eq. (5). Thus, we derive a theoretical LTNSTD to fit the score data for each of these samples. We can derive a series of theoretical LTNSTDs to fit these score data for each sample for a given f from 6 to 20. For these derived LTNSTDs, the Kolmogorov–Smirnov (KS) test [22] is then used to verify the goodness of fit between a theoretical cumulative distribution function and an empirical distribution function for each random sample. As a result, we are able to choose a well-fitted LTNSTD as a theoretical distribution for each random sample.

5. Third step of our data mining: discoveries of UFRs

For a continuous distribution of a random variable x, we define the quantile [24] q_α with probability α in the distribution as $P(x \le q_\alpha) = \alpha$. For a given probability α in the derived theoretical cumulative distribution, $F(x; f, \delta)$ of



Theoretical curves with f=8 are drawn with a jagged line and f=10 with a smooth line

Fig. 1. Empirical probability density functions (rows 1 and 3) and empirical distribution functions (rows 2 and 4) plotted together with linearly transformed theoretical probability density functions (rows 1 and 3) and cumulative distribution functions (rows 2 and 4) of *non-central t distribution*. The SIGSCR data computed in *M. gen* are shown in rows 1 and 2 and STBSCR data are shown in rows 3 and 4. The two scores computed by sliding the fixed windows of 100, 300 and 500 bases are shown in the left, middle and right, respectively. The empirical step functions are plotted with step size 0.2, and the theoretical curves with the degree of freedom $f = 8$ are drawn with a jagged line, and those with $f = 10$ are drawn with a smooth line.

Table 2

Analysis of variance for significance score (SIGSCR) and stability score (STBSCR) computed by sliding the window of 100, 300 and 500 nt along the complete *Mycoplasma genitalium* (*M. gen*) genome (strain G37)

Region	(a)Segment counts and the means of SIGSCR and STBSCR in *M. gen* genome								
	Window of 100 nt			Window of 300 nt			Window of 500 nt		
	N (Counts)	SIGSCR	STBSCR	*N* (Counts)	SIGSCR	STBSCR	*N* (Counts)	SIGSCR	STBSCR
Noncoding	35,214	−0.744	−0.151	25,513	−1.431	−0.532	21,317	−1.677	−0.702
Protein	501,941	−0.575	−0.024	45,8017	−1.348	0.057	416,168	−1.970	0.071
rRNA	4350	−0.792	−1.707	3655	−1.258	−2.276	3255	−1.421	−2.554
Source of variation	D.F.	Sum of squares		Mean squares		F-value and probability (tail)			
		SIGSCR	STBSCR	SIGSCR	STBSCR	SIGSCR	STBSCR		
(b) ANOVA table for the two scores distributed in *M. gen* genome (data from window of 100 nt)									
Between groups	2	1126	13,772	563	6886	457.2 (0.0000)	7314 (0.0000)		
Within groups	541,502	666,697	509,770	1.2312	0.9414				
(c) ANOVA table for the two scores distributed in *M. gen* genome (data from window of 300 nt)									
Between groups	2	201.6	27,603	100.8	13,801	63.38 (0.0000)	15,662 (0.0000)		
Within groups	487,182	774,863	429,305	1.5905	0.8812				
(d) ANOVA table for the two scores distributed in *M. gen* genome (data from window of 500 nt)									
Between groups	2	2657	33,764	1328	16,882	593.9 (0.0000)	19,626 (0.0000)		
Within groups	440,737	985,664	379,122	2.2363	0.8602				

Table 3
Extremes of unusual folding regions (UFRs) detected in *M. gen* (G37) genome. (Eight types of UFR (types 1–8) are defined based on the computed SIGSCR and/or STBSCR as listed in the second and third columns. Numbers listed in parentheses indicate the probabilities of the eight types of UFR occurring in the complete *M. gen* sequence according to the derived linearly transformed theoretical *non-central t distributions*. Numbers listed in the right three columns are the counts of these distinct UFRs that are involved entirely within the protein coding, RNA gene, and non-coding regions in the genome)

Type	SIGSCR (*P*-value)	STBSCR (*P*-value)	Protein	RNA	Non-coding
(a) UFR counts detected by sliding a window of 100 nt along *M. gen.* (G37)					
1	≤ -4.75 (0.0025)		547	2	275
2	≥ 2.24 (0.0025)		370	–	99
3		≤ -3.74 (0.0025)	397	284	111
4		≥ 2.48 (0.0025)	504	–	92
5	≤ -4.75 (0.0025)	≤ -2.81 (0.01)	314	2	111
6	≥ 2.24 (0.0025)	≥ 2.05 (0.01)	203	–	23
7	≤ -3.73 (0.01)	≤ -3.74 (0.0025)	251	34	53
8	≥ 1.74 (0.01)	≥ 2.48 (0.0025)	164	–	33
(b) UFR counts detected by sliding a window of 300 nt along *M. gen* (G37)					
1	≤ -5.91 (0.0025)		1048	–	279
2	≥ 1.94 (0.0025)		458	–	44
3		≤ -3.96 (0.0025)	151	168	42
4		≥ 2.38 (0.0025)	703	–	8
5	≤ -5.91 (0.0025)	≤ -2.91 (0.01)	107	–	84
6	≥ 1.94 (0.0025)	≥ 1.98 (0.01)	208	–	–
7	≤ -4.89 (0.01)	≤ -3.96 (0.0025)	99	12	20
8	≥ 1.37 (0.01)	≥ 2.38 (0.0025)	273	–	–

LTNSTD we calculate the quantile q_α by solving the equation $q_\alpha = F^{-1}(x; f, \delta)$, where $F^{-1}(x; f, \delta)$ is the inverse function of $F(x; f, \delta)$. In practice, the q_α is computed by the function NCTINV in the statistical toolbox of MATLAB software. In general, we calculate quantile, q_α, with probability $\alpha = 0.01$, 0.005, 0.0025 and 0.001 in the derived LTNSTD. For a desired value of α (a very small value), we can search for those UFRs in the sequence whose SIGSCR and/or STBSCR values are greater or less than the selected q_α. In this study, we define eight types of UFR termed types 1–8. For example, if SIGSCR ≤ -4.75, then the local segment of 100 bases is defined as the UFR of type 1. If STBSCR ≤ -3.74 the local segment of 100 bases is defined as the UFR of type 2. The probabilities α of the type 1 and 2 UFRs occurring in the *M. gen* genome are less than or equal to 0.0025 by chance.

6. Results and discussion

6.1. Statistics of SIGSCR and STBSCR in the M. gen genome

Statistics of local thermodynamic stability in the *M. gen* sequence are listed in Table 1. It is clear that the distributions of SIGSCR and STBSCR computed by windows of 100, 300 and 500 bases are asymmetric in the *M. gen* sequence. These distributions do not follow a normal distribution because of large skewness in the samples (see Fig. 1). We also computed the means of SIGSCR and STBSCR in the protein coding, RNA gene and non coding regions by means of the known gene structures of *M. gen* listed in Genbank (see Table 2a). The means of STBSCR in the domain of RNA genes computed by windows of 100, 300 and 500 bases were -1.707, -2.276 and -2.554, respectively. On average, the RNA gene domain was the most thermodynamically stable region among the three different domains. We also observed that the protein coding sequence was the least stable on average. Analyses of variance (ANOVA) for these data indicate that the thermodynamic stability of the local segment within the three different domains is remarkably different from each other (see Table 2b–d). The observed bias toward more thermodynamically stable folding segments in the RNA genes is very statistically significant by ANOVA test.

6.2. Non-central t distributions of SIGSCR and STBSCR in M. gen

Our data indicate that the errors between the derived theoretical cumulative distribution function and empirical distribution function are not sensitive to f values for the random samples. The derived two LTNSTDs of data SIGSCR and STBSCR show the acceptance limit with the significance level 0.05 or above for f from 10 to 20 for the two scores computed by a window of 100 bases. For other samples, we obtained similar results. The linearly transformed theoretical probability density functions and cumulative distribution functions of *non-central t distributions* for SIGSCR and STBSCR derived by $f = 8$ and 10 are shown in Fig. 1. In the plots, the empirical distribution functions are all fitted well using the derived cumulative distribution functions. Our results show that the LTNSTD is a good statistical model to describe the distribution of SIGSCR and STBSCR computed in the *M. gen* sequence.

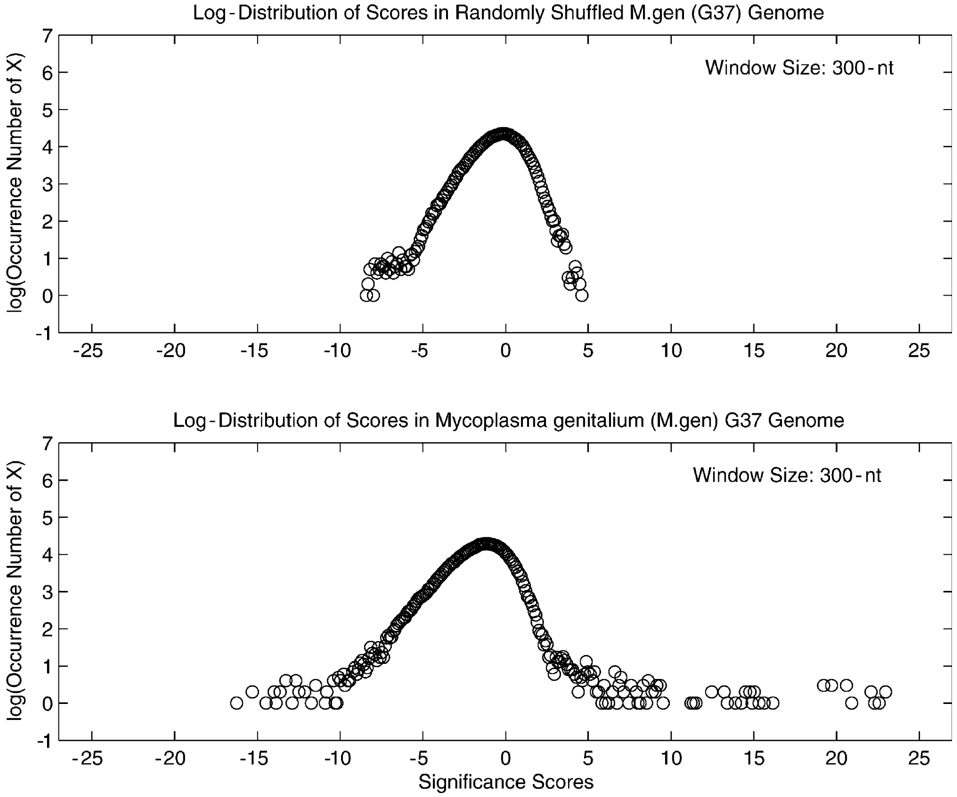


Fig. 2. Significance score (SIGSCR) of thermodynamic stability for the local segment of 300 bases computed in the artificial (top) and native *M. gen* genome (bottom). The artificial genome sequence was produced by randomly shuffling the complete *M. gen* genome. The horizontal axis represents SIGSCR and vertical axis represents the log-scale of the occurrence number (base 10 logarithm) of each SIGSCR in the plot.

6.3. Statistical extremes of UFR in the sequence

Based on values of SIGSCR and STBSCR, we discover eight types of UFR (types 1–8) in the *M. gen* genome (see Table 3). Among them, type 5 UFR is defined if its SIGSCR is less than or equal to −4.75 and its STBSCR is less than or equal to −2.81. Thus, the type 5 UFR represents the segment sequence that is significantly more stable than both their randomly shuffled sequences and other fragments of the same size in the complete genome. The extremely stable type 5 UFRs often provide useful information for searching for biologically functional elements in the bacteriophage [9] and eukaryotic mRNAs [5,7,8,10–12]. The UFRs detected in the *M. gen* sequence by the window sizes of 100 and 300 bases are summarized in Table 3. The detailed data about these detected UFRs are available on request from the authors.

To obtain a clear idea about what difference of the statistical extremes are selected in the native genome and its corresponding, randomly shuffled genome sequence, we applied the same approach to the artificial sequence of the randomly shuffled genome using a fixed window of 300 bases. The comparisons of the SIGSCR data computed in the native and artificial samples are shown in Figs. 2 and 3. Our results indicate that there is not a simple relation between SIGSCR and the percentage of base G + C in the fragment. We consider that SIGSCR in *M. gen* is strongly dependent on the distinct sequence pattern in the local segment. It also indicates that the range of the SIGSCR value computed in the artificial sequence is from −8.42 to 4.61. Among them, only 18 out of the 579,775 segments have SIGSCR that is less than or equal to −7.75. However, we detected 241 such extreme UFRs in the native *M. gen* genome. The probability of these extreme UFRs in *M. gen* is about 0.0005 or less. It is clear that the UFRs extracted from the natural genome are not selected by random. It may imply some biological functions involved in the distinct sequence pattern, where the folded structure of RNA or DNA segments plays an important role in their functions.

The program SIGSTB was implemented in Fortran 77 on a Silicon Graphics (SGI) Computer with IRIX 6.5. It has also been executed on a Compaq/DEC Alpha 8400/625 EV56 with Digital Unix. Our method for calculating SIGSCR and STBSCR requires $O(w^3N)$ computation time, where w is the window size and N is the sequence length. For example, it took 18,788 CPU seconds on a SGI Octane computer for calculating two scores in the sequence 1–300,600 of *M. gen*using a fixed window of 300 bases. All statistical analyses in this study were performed using the Statistical Toolbox of MATLAB software package (http://www.mathworks.com).

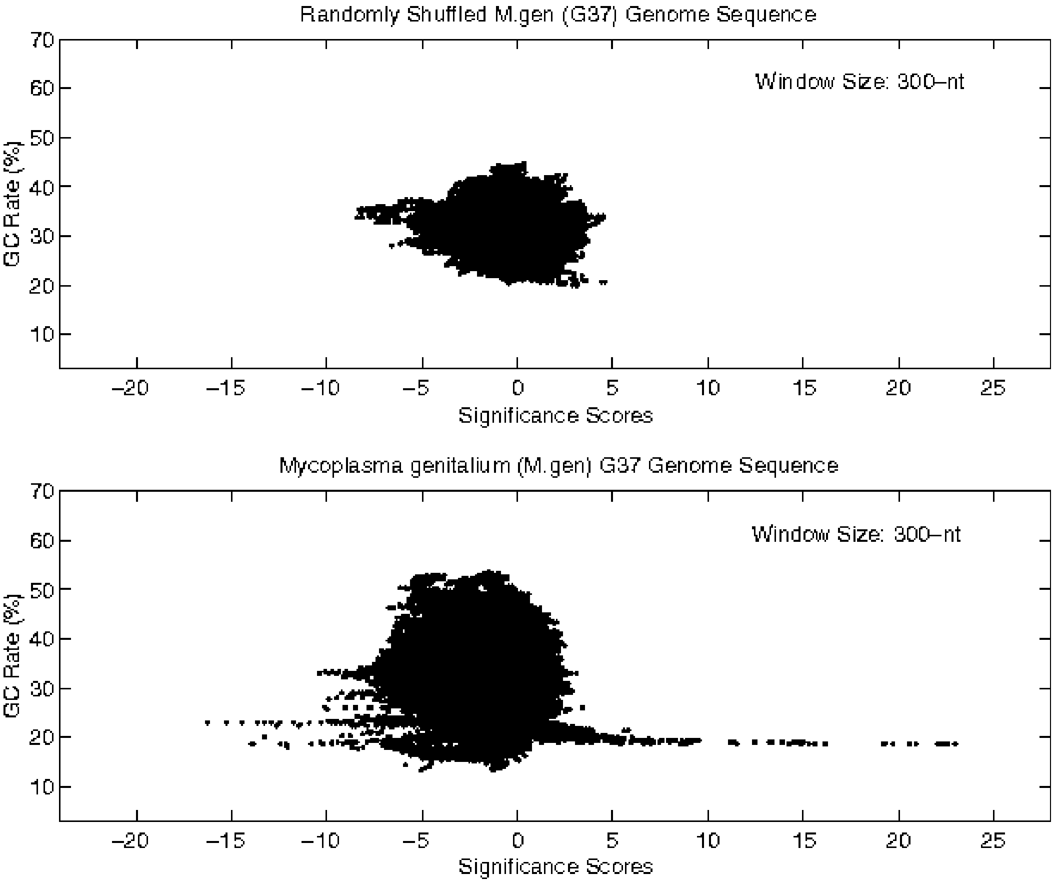


Fig. 3. Relationships between SIGSCR and base compositions of G + C computed from the local segment of 300 bases in the artificial (top) and natural *M. gen* genome (bottom). The horizontal axis represents SIGSCR and vertical axis represents the percentage of base G + C in the local segment. The SIGSCR and percentage of base G + C were computed by sliding the fixed window of 300 bases stepped one base at a time along the sequences.

7. Conclusions and perspectives

In this study, we present a data mining approach to discover UFRs in the *M. gen* genome sequence. At the first stage of the approach, we calculate two *z*-scores of SIGSCR and STBSCR in the sequence. Next, we derive a LTNSTD statistical model to describe the distributions of the two scores in the *M. gen* sequence. Finally, we discover the UFRs in *M. gen* based on the derived LTNSTDs, whose SIGSCR and STBSCR values are significantly deviated from their sample means. The approach is generally applicable for other genomes. For instance, we also computed the two scores in other microbial genomes, such as *Helicobacter pylori* strains 26695 and J99, and *Mycoplasma pneumonia*. The distributions of the two scores in these sequences are well represented by a LTNSTD. Statistical extremes of UFRs can be confidently assessed based on the derived, theoretical LTNSTD. The precise locations for these UFRs can be further inferred by an extended search (SEGFOLD) in which the window size is systematically changed in the corresponding extended regions. These detected UFRs in *M. gen* and others can be suggested as candidate sites for further experimental study in searching gene regulatory elements and potential target sequences of long-chain antisense RNAs. Our data mining approach in the genomic sequence is particularly useful for antisense RNA therapeutics and the targeting of RNA-binding drugs against pathogenic bacteria.

Acknowledgements

The contents of this publication do not necessarily reflect the views or policies of the Department of Health and Human Services, neither does mention of trade names, commercial products, or organizations imply endorsement by the US Government. The program SEGFOLD and its modified version SIGSTB are available via anonymous ftp as /home/ftp/pub/users/shuyun/sigfold at ftp.ncifcrf.gov. The script files of performing these statistical analyses in this study are available upon request from the authors.

References

[1] S. Karlin, A.M. Campbell, J. Mrazek, Comparative DNA analysis across diverse genomes, Annu. Rev. Genet. 32 (1998) 185–225.

[2] R. Durbin, S. Eddy, A. Krough, G. Mitchison, Biological Sequence Analysis, Cambridge University Press, Cambridge, UK, 1998.

[3] The RNA World, in: R.F. Gesteland, T.R. Cech, J.F. Atkins (Eds.), 2nd ed., Cold Spring Harbor Lab. Press, New York, 1999.

[4] RNA Structure and Function, in: R.W. Simons, M. Grunberg-Manago (Eds.), Cold Spring Harbor Lab. Press, New York, 1998.

[5] M.H. Malim, J. Hauber, S.-Y. Le, J.V. Maizel Jr., B.R. Cullen, The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA, Nature 338 (1989) 254–257.

[6] S.-Y. Le, M.H. Malim, B.R. Cullen, J.V. Maizel Jr., A. highly, conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses, Nucl. Acids Res. 18 (1990) 1613–1623.

[7] S.-Y. Le, J.-H. Chen, M.J. Braun, M.A. Gonda, J.V. Maizel Jr., Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-1), Nucl. Acids Res. 16 (1988) 5153–5168.

[8] S.Y. Le, J.-H. Chen, J.V. Maizel Jr., Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retro-viruses, Nucl. Acids Res. 17 (1989) 6143–6152.

[9] S.Y. Le, J.-H. Chen, J.V. Maizel Jr., Identification of unusual RNA folding patterns encoded by bacteriophage T4 gene 60, Gene 124 (1993) 21–28.

[10] S.Y. Le, A. Siddiqui, J.V. Maizel Jr., A common structural core in the internal ribosome entry sites of picornavirus, hepatitis C virus, and pestivirus, Virus Gene 12 (1996) 135–147.

[11] O. Sella, G. Gerlitz, S.-Y. Le, O. Elroy-Stein, Differentiation-induced internal translation of c-*sis* mRNA: analysis of the *sis* elements and their differentiation-linked binding to the hnRNP C protein, Mol. Cell. Biol. 19 (1999) 5429–5440.

[12] S. Chen, S.-Y. Le, D.L. Newton, J.V. Maizel Jr., S.M. Rybak, A gender-specific mRNA encoding a cytotoxic ribonuclease contains a $3'$ UTR of unusual length and structure, Nucl. Acids Res. 28 (2000) 2375–2382.

[13] S.-Y. Le, J.V. Maizel Jr., A method for assessing the statistical significance of RNA folding, J. Theor. Biol. 138 (1989) 495–510.

[14] S.-Y. Le, J.-H. Chen, J.V. Maizel Jr., Efficient searches for unusual folding regions in RNA sequences, in: R.H. Sharma, M.H. Sharma (Eds.), Structure and Methods: Human Genome Initiative and DNA Recombination, vol. 1, Adenine Press, Schenectady, 1990, pp. 127–136.

[15] M. Evans, N. Hastings, B. Peacock (Eds.), Statistical Distributions 2nd ed., Wiley, New York, 1993.

[16] Testing Statistical Hypothesis, in: E.L. Lehmann (Ed.), Wiley, New York, 1959.

[17] M. Zuker, Predication of RNA secondary structure by energy minimization, Methods Mol. Biol. 25 (1994) 267–294.

[18] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, D.H. Turner, Improved free-energy parameters for predictions of RNA duplex stability, Proc. Natl. Acad. Sci. USA 83 (1986) 9373–9377.

[19] M. Elloumi, New algorithms to predict secondary structures of RNA macromolecules, in: Proceedings of 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE'98, vol. 1, Benicassim, Spain, Springer Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, 1998, pp. 864–875.

[20] M. Elloumi, Algorithms for the prediction of secondary structures of RNA macromoles, Egyptian Computer Journal, Institute of Statistical Studies and Research, Cairo University, Egypt, 2000, 378–387.

[21] J.-H. Chen, S.-Y. Le, B. Shapiro, K.M. Currey, J.V. Maizel Jr., A computational procedure for assessing the significance of RNA secondary structure, CABIOS 6 (1990) 7–18.

[22] R.V. Hogg, E.A. Tanis (Eds.), Probability and Statistical Inference 5th ed., Prentice-Hall, Upper Saddle River, NJ, 1997.

[23] S.-Y. Le, W.-M. Liu, J.-H. Chen, J.V. Maizel Jr., Local thermodynamic stability scores are well represented by a non-central Student's *t* distribution, J. Theor. Biol. 210 (2001) 411–423.

[24] R.V. Hogg, A.T. Craig (Eds.), Introduction to Mathematical Statistics, 5th ed., Prentice-Hall, Upper Saddle River, NJ, 1995.